Revista Română de Interacțiune Om-Calculator 2 (2009), 1-17

Resurse lingvistice pentru un sistem de întrebare-răspuns pentru limba română

Verginica Barbu Mititelu, Alexandru Ceauşu, Radu Ion,
Elena Irimia, Dan Ştefănescu, Dan Tufiş

Institutul de Cercetări pentru Inteligență Artificială, Academia Română
Calea 13 Septembrie, nr. 13, 050711, Bucureşti
E-mail: {vergi, aceausu, radu, elena, danstef, tufis}@racai.ro

Rezumat. În acest articol descriem câteva resurse lingvistice (un lexicon, o morfologie paradigmatică, două tezaure lingvistice – wordnetul românesc şi Eurovoc – şi un corpus multilingv paralel) din perspectiva utilității lor preponderent în sistemele de întrebare-răspuns. Prezentăm aici etapele procesului de găsire automată a unui răspuns la o întrebare formulată în limbaj natural de utilizator. Acolo unde se impune, este înfățişat modul în care resursele lingvistice contribuie la rezolvarea diverselor probleme. Lexiconul funcționează ca un corector ortografic pentru întrebarea introdusă de utilizator. Morfologia paradigmatică este utilă în procesul de lematizare a întrebării şi a corpusului. Wordnetul românesc îşi dovedeşte eficacitatea în expandarea frazei de interogare, în identificarea lanțurilor lexicale între sensurile cuvintelor şi în găsirea răspunsurilor într-un sistem mono- şi multilingv. Tezaurul Eurovoc este folosit la segmentarea şi lematizarea întrebării introduse de utilizator şi a corpusului multilingv paralel în care se caută răspunsul la această întrebare. Arhitectura sistemului de întrebare-răspuns descris aici nu este dependentă de limbă; resursele lingvistice, însă, sunt, în mod intrinsec, specifice unei limbi (vezi lexiconul şi morfologia paradigmatică); excepție fac cele a căror organizare sau structură permite o perspectivă multilingvă (ex. tezaurele şi corpusul), în cazul nostru ele fiind aliniate între ele.

Cuvinte cheie: sistem de întrebare-răspuns, resurse lingvistice, tezaur, ontologie lexicală, corpus.

1. Introducere

Interesul pentru lingvistica computațională şi pentru finanțarea proiectelor din acest domeniu de cercetare este susținut de aplicațiile pe care le dezvoltă, cu impact pentru societate în ansamblul ei. Lumea de astăzi are, prin internet, acces la o cantitate de informații extrem de vastă. Sunt necesare instrumente care să permită găsirea datelor petinente în acest volum uriaş de cunoştințe. Comunitatea lingvisticii computaționale ține

pasul cu aceste nevoi şi încearcă să ofere sisteme informatice menite să ajute utilizatorul să se descurce în noianul de documente disponibile.

Pentru informare, utilizatorul recurge frecvent la formularea unei întrebări. Pornind de la aceasta, diverse aplicații oferă variate soluții: documentele care, se presupune, includ răspunsul la întrebare sau chiar fragmentele considerate adecvate din documentele estimate a fi cele mai relevante.

Regăsirea informației (engl. *information retrieval*) este știința aspectelor teoretice şi practice ale căutării informațiilor de interes în colecții mari de date, ale creării şi menținerii acestor colecții (Jackson şi Moulinier, 2002 : 26).

Extragerea informației (engl. *information extraction*) este un tip de regăsire a informației, care-şi propune să identifice, în mod automat, anumite tipuri de entități, relații, evenimente în texte. Se deosebeşte de regăsirea informației prin faptul că nu îşi propune găsirea de documente, ci de informație relevantă în interiorul documentelor (Jackson şi Moulinier, 2002 :75).

Sistemele de întrebare-răspuns (engl. *question answering*) găsesc, în mod automat, şi redau răspunsuri la întrebări formulate în limbaj natural (Harabagiu şi Moldovan, 2003 : 561). De la utilizatori ai Internetului dornici să-şi îmbogățească cunoştințele generale, la analişti specialişti într-un domeniu, care vor să afle o informație specifică, pentru toți astfel de sisteme devin necesare şi de neînlocuit.

2. Utilitatea resurselor lingvistice în aplicațiile lingvisticii computaționale

Aplicațiile de regăsire şi extragere a informației şi cele de întrebare-răspuns au la dispoziție doar cuvintele introduse de utilizator în întrebare. În primul rând, trebuie verificată ortografia acestora, iar pentru aceasta este util un *lexicon*. În al doilea rând, cuvintele introduse de utilizator pot să nu se regăsească în documente exact în forma din întrebare. În astfel de cazuri îşi dovedeşte utilitatea un *lematizor*. În al treilea rând, cuvintele folosite de utilizator nu sunt neapărat cele mai relevante pentru găsirea răspunsului. De aceea, este necesară expandarea frazei de interogare, prin luarea în considerare a cuvintelor aflate în relații semantice sugestive (sinonimie, hiponimie, hiperonimie) cu cele care compun această frază de interogare.

Pentru aceasta este nevoie de un *tezaur*, în care sunt înregistrate cuvinte cu sensurile lor şi cu relaţiile lexico-semantice existente între ele. În al patrulea, dar în nici un caz ultimul rând, este necesară o colecţie de documente în care să fie căutat răspunsul. Această colecţie poate fi un *corpus* sau internetul.

În acest articol, prezentăm, din perspectiva utilităţii (preponderent) în sistemele de întrebare-răspuns, cinci resurse lingvistice disponibile la Institutul de Cercetări pentru Inteligenţă Artificială (ICIA) al Academiei Române: un lexicon, o morfologie paradigmatică pentru limba română, un tezaur lingvistic general (ontologie lexicală), un tezaur specializat şi un corpus de mari dimensiuni.

2.1 Lexicon

Motoarele de căutare iau cuvintele din întrebarea introdusă de utilizator şi întorc rezultate ordonate pentru aceasta. În cadrul acestui proces, un factor important care duce la rezultate nesatisfăcătoare este incorectitudinea scrierii cuvintelor de către utilizator. Se pare că 10-15% dintre frazele de căutare conţin astfel de greşeli (Cucerzan şi Brill, 2004). De multe ori este vorba despre erori dactilografice, însă cele mai dăunătoare sunt cele rezultate din confuzia unor termeni: paronimia (i.e. relaţia dintre două cuvinte asemănătoare din punctul de vedere al formei, dar deosebite ca sens), de pildă, poate crea confuzii numeroase: *revela* şi *releva*. În funcţie de colecţia de documente în care se face căutarea, se preferă diverse modalităţi de corectare a acestor erori. În cazul nostru, al căutării într-un corpus[1] îşi poate dovedi utilitatea un lexicon cuprinzător. Fiecare element din întrebare (supusă în prealabil unei operaţii de segmentare la nivel de cuvânt) este verificat din punct de vedere ortografic prin căutarea formei în lexicon. Dacă aceasta nu este găsită, se aplică o euristică de găsire a celei mai apropiate forme pe baza unor principii simple: erori de dactilografiere

[1] Pentru căutarea răspunsurilor pe web, se impun alte modalităţi de corectare a erorilor de scriere, determinate de vastitatea materialului lexical ce trebuie acoperit şi de raritatea utilizării anumitor cuvinte (vezi Chen et al. 2007).

(inversarea unor litere apropiate pe tastatură), litere alăturate în cuvânt inversate, litere duble.

Lexiconul românesc a fost creat în cadrul proiectului MULTEXT-EAST şi conţine forme flexionare ale cuvintelor din limba română, împreună cu descrierea lor morfosintactică şi cu lema (i.e. forma de dicţionar, Tufiş et al., 1997, vezi Figura 1). După încheierea proiectului, în ICIA s-a continuat îmbogăţirea acestui lexicon, iar în momentul de faţă, el conţine aproape 1.223.000 de intrări (i.e. forme flexionare), faţă de 450.000 câte conţinea la finalizarea MULTEXT-EAST (cf. Tufiş, 2002).

Formă ocurentă	Lemă	MSD	CTAG
aramă	aramă	Ncfsrn	NSRN
arama	aramă	Ncfsry	NSRY
arame	aramă	Ncfp-n	NPN
arame	aramă	Ncfson	NSON
aramei	aramă	Ncfsoy	NSOY
aramele	aramă	Ncfpry	NPRY
aramelor	aramă	Ncfpoy	NPOY

Figura 1. Formele flexionare ale substantivului *aramă* în lexicon.

O situaţie aparte care impune corectură grafică o reprezintă introducerea de către utilizator a cuvintelor fără diacritice sau numai parţial cu diacritice. Cum corpusul în care motorul nostru face căutarea (vezi *infra* 2.4) conţine diacritice, se impune introducerea acestora şi în întrebare. În cadrul ICIA a fost implementat sistemul DIAC (Tufiş şi Chiţu, 1999) extins şi mult îmbunătăţit de sistemul DIAC^{+}, care are o precizie de 97.75% în reconstituirea diacriticelor în texte neprocesate (Tufiş şi Ceauşu, 2008).

Principala problemă ridicată de utilizarea lexiconului în corectarea grafiei o reprezintă limitarea cantitativă a acestuia. În consecinţă, inexistenţa unui cuvânt din întrebare în lexicon nu trebuie interpretată neapărat ca greşeală de scriere.

2.2 Morfologie paradigmatică pentru limba română

Odată stabilită corectitudinea grafică a întrebării utilizatorului urmează adnotarea de tip stratificat (Tufiş, 1999) cu etichete morfosintactice (vezi Ion, 2007 pentru descrierea completă a acestui proces) şi lematizarea ei,

adică identificarea formei de bază (lema) a fiecărui cuvânt, cu ajutorul lexiconului. Dată fiind o formă ocurentă şi descrierea sa morfosintactică, se regăseşte uşor lema formei respective în lexiconul de forme flexionate. Pentru cazurile neacoperite de acesta, la ICIA a fost dezvoltat un lematizor (Ion, 2007: 22-26) care foloseşte un set de reguli (specific fiecărei categorii gramaticale flexionare) induse automat din lexicon, care generează leme candidat pentru cuvântul necunoscut, şi apoi modele Markov (antrenate pe leme din lexicon) pentru a ordona candidații. Candidatul cu cea mai mare probabilitate câştigă. Procedura funcționează foarte bine, cele mai multe dintre erori sunt în cazul cuvintelor necunoscute care aparțin paradigmelor flexionare neregulate sau atunci când adnotarea morfosintactică a formei ocurență a fost greşită. În ansamblu, ținând cont de vasta acoperire a lexiconului şi de rata mică de eroare a lematizorului statistic, probabilitatea unei erori de lematizare este neglijabilă.

În vederea căutării răspunsului la întrebarea utilizatorului, se impune şi preprocesarea corpusului. El este supus, pe rând, operațiilor de segmentare, adnotare cu etichete morfosintactice şi lematizare. Această ultimă etapă se face tot cu ajutorul lematizorului descris mai sus. Pentru formele inexistente în lexicon, însă, mai dispunem de o modalitate de identificare a lemei, cu ajutorul morfologiei paradigmatice pentru limba română (ROPMORPH, Tufiş, 1989, Irimia, 2009a), care conține un inventar complet al terminațiilor din limba română pentru verbe, substantive şi adjective, fiecare terminație fiind asociată informației morfologice caracteristice unei anumite forme (vezi Figura 2).

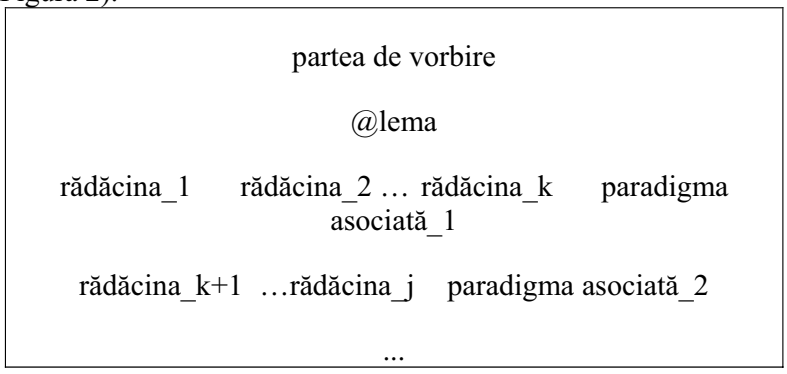

Figura 2. Structura unei intrări în ROPMORPH.

Astfel, pentru fiecare formă flexionară, poate fi obținută o listă de rădăcini prin eliminarea terminațiilor specifice fiecărei paradigme aplicabile formei. În cazul în care lista conține un singur element, cu ajutorul generatorului morfologic (ROG) (Irimia, 2009b) se poate, foarte uşor, genera lema corespunzătoare formei flexionare de la care am pornit. În cazul în care lista conține două sau mai multe rădăcini, Irimia (2009a) descrie un algoritm prin care ROG generează toate formele posibile din toate paradigmele asociate acestor rădăcini, apoi apelează o serie de funcții dintr-o bibliotecă dezvoltată la ICIA, care invocă motorul de căutare Google™ şi extrag informații cantitative, pentru fiecare cuvânt din fiecare paradigmă, referitoare la numărul de apariții pe web. Sunt eliminate toate formele care apar de mai puțin de 10 ori (se presupune că ele se datorează unor greşeli de scriere), iar fiecare paradigmă este ierarhizată în funcție de suma ocurențelor pe web ale formelor sale. Este aleasă paradigma cu cel mai mare scor, iar apoi ROG generează lema.

2.3. Resurse lexico-semantice

Tezaurul este un tip de dicționar în care cuvintele cu sensuri asemănătoare sunt grupate la un loc (cf. Online Cambridge Dictionary). Prima lucrare de acest fel, din epoca modernă, a fost creată de Peter Mark Roget în 1805 şi cuprindea 15.000 de cuvinte. De atunci a cunoscut adăugiri succesive, la început de către urmaşi ai lui Roget, apoi de către alți specialişti, rezultatul fiind câte o ediție nouă a acestei lucrări, ultima dintre ele (din 1987) cuprinzând peste 1.250.000 de cuvinte.

Organizarea materialului lingvistic se face în clase, fiecare având o ierarhie internă, asemănătoare unui arbore cu multiple ramificații pentru grupuri de sensuri. Scopul unei asemenea lucrări este să faciliteze găsirea cuvintelor înrudite semantic (sinonime, antonime, hiperonime, hiponime), ceea ce în dicționarele explicative este mai greu sau chiar imposibil, şi să ajute la alegerea celui mai potrivit cuvânt pentru exprimarea sensului dorit de utilizator.

În prelucrarea limbajului natural se folosesc în ultimii ani, din ce în ce mai mult, ontologiile lexicale. Ca şi în tezaurele binecunoscute în regăsirea de documente, în ontologiile lexicale sunt stocate numeroase cuvinte (în forma lor lematizată), împreună cu relațiile semantice sau lexicale dintre ele. Cuvintele sunt abstractizate ca unități semantice pentru exprimarea

conceptelor. Între ele se stabilesc relații ierarhice (pentru indicarea termenilor mai generali şi a celor mai specifici), de echivalență (stabilite între sinonime) sau asociative (relații care nu sunt nici ierarhice, nici de echivalență, dar se stabilesc între cuvinte, iar utilizatorul le acceptă ca răspuns la întrebarea sa). Ceea ce diferențiază o ontologie lexicală de un tezaur este nivelul axiomatic al descrierii conceptelor denotate de cuvintele reprezentate. Pe lângă definiție, relațiile de tip ierarhic (hiperonim, hiponim), meronimic sau de altă natură, o ontologie lexicală pune în evidență o legatură de tip ontologic, la un concept interlingual caracterizat de o serie de definiții şi axiome. De exemplu, pentru cuvântul „container", descrierea în ontologia SUMO/MILO (Niles şi Pease, 2001) este prezentată în Figura 3.

Figura 3: Descrierea axiomatică a conceptului interlingual „container"

2.3.1 RO-Wordnet, ontologie lingvistică pentru limba română

Lexicul este, fără îndoială, cea mai importantă resursă lingvistică a unei limbi. Marea majoritate a cercetărilor actuale, atât în lingvistica formală, cât mai ales în tehnologia limbajului, plasează componenta lexicală în centrul

modelelor de limbă, sub influenţa a ceea ce a fost numită abordarea *lexicalizată* sau *lexicalistă* a studiului limbii. Studiul computaţional al dicţionarelor electronice, natura informaţiei ce trebuie inclusă în ele şi tipul de prelucrări pe care le poate facilita o anumită structurare a unui mare volum lexical a fost, fără îndoială, fundamental influenţat de proiectul WordNet, lansat în 1985 la Universitatea din Princeton sub conducerea reputatului psiholingvist George Miller (Fellbaum, 1998). WordNet, resursă publică, în varianta actuală, este o uriaşă reţea semantică lexicală, în care peste 117.000 de *înţelesuri* lexicalizate în limba engleză prin mai mult de 155.000 de cuvinte sunt asociate între ele prin relaţii semantice (Fellbaum, 1998). Fondul lexical este distribuit în patru reţele semantice corespunzând categoriilor gramaticale deschise: substantive, verbe, adjective şi adverbe. Noţiunea de *înţeles* (engl. *meaning*) este echivalată în WordNet cu cea de *concept* şi este reprezentată printr-o serie sinonimică în care fiecărui cuvânt al seriei îi este asociat un număr ce identifică sensul în care cuvântul respectiv are înţelesul asociat conceptului. Seria sinonimică ce identifică un înţeles se numeşte *sinset* (engl. *syn(onymy)set*). Relaţiile existente între *sinseturi* sunt de diferite tipuri, depinzând de categoria gramaticală a cuvintelor ce alcătuiesc un anumit sinset (antonimie/sinonimie, hiponimie/hiperonimie, holonimie/meronimie, troponimie etc.) şi având justificare psiholingvistică.

Wordnetul este un dicţionar semantic care rafinează structurările întâlnite în tezaure, prin considerarea aspectelor lexicale ale materialului lingvistic conţinut: de exemplu, reţeaua semantică a substantivelor cuprinde 25 de structuri arborescente (relaţiile organizatoare fiind hiponimia şi meronimia), iar cea a verbelor 15 astfel de structuri (organizate în funcţie de hiponimie şi troponimie), asemănătoare claselor din *Tezaurul* lui Roget. Adjectivele şi adverbele nu se pretează, însă, la astfel de organizări ierarhice: în cazul adjectivelor se preferă gruparea lor în funcţie de relaţia de antonimie, iar adverbele (cu excepţia cazurilor rare de antonimie) nu au nicio organizare.

Influenţa proiectului WordNet a fost enormă în domeniul tehnologiei limbajului (exprimată poate şi prin faptul că acum, în limbajul tehnic cel puţin, cuvintele „wordnet” şi „synset” au devenit substantive comune, importate prin calchiere în mai toate limbile), iar beneficiile acestui concept atât de evidente încât Comisia Europeană a decis finanţarea a două proiecte de mare anvergură: EuroWordNet (Bloksma et al., 1996) şi BalkaNet (Stamou et al., 2001).

În cadrul BalkaNet a început dezvoltarea (Tufiş şi Barbu, 2004) ontologiei lingvistice pentru limba română (Ro-WN), păstrându-se organizarea materialului lingvistic din wordnetul englezesc (Tufiş, 2004). Consolidarea acestei resurse lingvistice a continuat în cadrul programului național „Cercetare de Excelență" finanțat de Ministerul Educației şi Cercetării prin proiectul ROTEL (http://www.ici.ro/~badea/rotel/index.htm) şi ca temă de plan a Institutului. În momentul scrierii acestei lucrări, Ro-WN conține peste 55.000 de sinseturi şi peste 47.000 de literali unici: 40.495 de sinseturi substantivale (adică 73%), 9.940 de sinseturi verbale (18%), 3831 de sinseturi adjectivale (7%) şi 834 de sinseturi adverbiale (2%). Toate sinseturile conținute în Ro-WN sunt aliniate cu echivalenții de traducere din limba engleză din Princeton WordNet (versiunea 2.0) şi, în plus, au legături ontologice la conceptele din ontologia de nivel superior şi mediu SUMO/MILO (vezi Figura 3). Sinseturile ontologiei lexicale pentru limba română beneficiază şi de marcaje de tipul clasificării zecimale universale (Bentivogli et al., 2004) şi pentru specificarea triadei de subiectivitate (Esuli şi Sebastiani, 2006) <obiectiv, pozitiv, negativ> (a se vedea pentru detalii Tufiş, 2008).

În cadrul unui sistem de întrebare-răspuns, o ontologie lexicală precum wordnetul îşi dovedeşte utilitatea în mai multe etape ale lucrului: expandarea frazei de interogare, identificarea lanțurilor lexicale între sensurile cuvintelor şi găsirea răspunsurilor într-un sistem monolingv sau multilingv.

2.3.1.1 Expandarea frazei de interogare

După prelucrare, întrebarea este transformată în frază de interogare în metalimbajul sistemului de căutare. De exemplu, întrebarea utilizatorului

„Care este scopul modificării deciziei 77/270/Euratom în ceea ce priveşte energia nucleară?"

devine

„label:REASON +("scop modificare decizie 77/270"^4 (scop AND modificare AND decizie AND 77/270) Euratom energie_nucleară "scop modificare decizie"^3 (scop AND modificare AND decizie) "modificare decizie 77/270"^3 (modificare AND decizie AND 77/270) "scop modificare"^2 (scop AND modificare) "modificare decizie"^2 (modificare AND decizie) "decizie 77/270"^2 (decizie AND 77/270) scop modificare

decizie 77/270 energie_nucleară)". Sistemul caută un paragraf etichetat ca fiind de tip „scop" (label:REASON), în care să existe „scop", „modificare", „decizie", „77/270" etc. Notaţiile de tip ^2 reprezintă un coeficient de creştere a relevanţei (engl. *boosting*) în căutare a termenului căruia i se aplică[2].

Întrebarea introdusă de utilizator poate prezenta deficienţe şi în ceea ce priveşte proprietatea termenilor utilizaţi (i.e. folosirea acelor cuvinte care să reflecte cel mai bine sensul). Din acest motiv, pentru găsirea răspunsului, este utilă expandarea frazei de interogare, prin includerea în căutare a cuvintelor sinonime şi chiar hiperonime, hiponime sau troponime cu cele introduse de utilizator. Organizarea ontologiei lexicale slujeşte foarte bine acestui scop: wordnetul conţine sinseturi (grupuri de sinonime) ierarhizate conform relaţiilor de hiponimie şi troponimie. Folosind un sistem de dezambiguizare automată a sensurilor cuvintelor din fraza de interogare (Ion şi Tufiş, 2007), se caută în Ro-WN cuvintele aflate în relaţii semantice relevante cu acestea (în special sinonimie, hipo- şi hiperonimie, troponimie) şi sunt luate în considerare în căutarea răspunsului în corpus.

2.3.1.2 Identificarea lanţurilor lexicale între sensurile cuvintelor

Pentru a putea ordona mulţimea răspunsurilor găsite de sistem pentru o frază de interogare, este nevoie să se stabilească un scor de apropiere semantică între cuvintele introduse de utilizator şi cele materializate în text (acestea nefiind întotdeauna identice, motiv pentru care se efectuează şi expandarea descrisă în 2.3.1.1). Calcularea acestui scor de apropiere semantică se face pe baza lungimii lanţurilor lexicale dintre cuvinte, folosind ontologia wordnet pentru limba română.

Un lanţ lexical reprezintă o mulţime de cuvinte înrudite semantic, specifice unui domeniu de discurs. Deosebirea faţă de cuvintele care formează un câmp semantic o reprezintă faptul că un lanţ lexical trebuie interpretat ca o cale în ierarhia Ro-WN, cale ce include, cel mai adesea, şi relaţii intercategoriale ce dau seama de dependenţele sintagmatice ce pot

[2] Pentru descrierea componentelor, a modului de funcţionare şi a ordonării rezultatelor sistemului de întrebare-răspuns dezvoltat în cadrul ICIA, vezi Ion et al. (2008).

exista între cuvintele care definesc capetele unui lanț lexical (vezi Ion et al., 2008) pentru descrierea acestei metode de ordonare a răspunsurilor). Diverse relații participă în mod diferit la stabilirea scorului de asemănare între două cuvinte: hiponimia şi relațiile derivative (intercategoriale) au cea mai importantă contribuție.

În exemplul de mai jos redăm trei lanțuri lexicale găsite între *fiu* şi *tată*, precedate de scorurile de asemănare (*near-antonym* notează relația de antonimie, *gloss* indică ocurența în definiție, *hypernym* notează relația de hiperonimie, iar *ss* relația de sinonimie, deci coocurența în acelaşi sinset):

```
0.333

     fiu(?) near_antonym fiică(1) <=> fiică(1) near_antonym copil(?)
gloss tată(?)

0.666

     fiu(?) hypernym băiat(1) <=> băiat(1) hypernym copil(?) gloss
tată(?)

0.733

     fiu(?) ss băiat(5) <=> băiat(5) ss copil(?) gloss tată(?)
```

2.3.1.3 Găsirea răspunsurilor într-un sistem de întrebare-răspuns multilingv

Imaginând un scenariu în care utilizatorul vrea să afle răspunsul într-o limbă străină (limba țintă pe care nu o stăpâneşte foarte bine) la o întrebare pe care preferă să o formuleze în limba maternă (limba sursă), se poate dovedi utilitatea resurselor lingvistice multilingve aliniate. Este vorba despre tezaure lingvistice sau ontologii lexicale multilingve aliniate între ele şi despre corpusuri aliniate la nivel de paragraf sau chiar de propoziție (vezi *infra* 2.4).

Procedura de lucru este identică, până într-un punct, cu cea a sistemului de întrebare-răspuns monolingv. Întrebarea este prelucrată, fraza de interogare este expandată şi se caută răspunsul în corpusul în limba sursă. Cum Acquis-ul comunitar (vezi *infra* 2.4 pentru detalii) în limba română este aliniat la nivel de paragraf cu Acquis-ul comunitar al tuturor țărilor membre ale Uniunii Europene, se poate identifica foarte rapid paragraful corespunzător din limba țintă (Ceauşu et al., 2006).

O altă modalitate de găsire a răspunsului într-o limbă țintă este de a traduce automat întrebarea utilizatorului şi de o a supune prelucrării, expandării şi apoi de a căuta răspunsul în corpusul din acea limbă. Această

metodă, însă, va da rezultate a căror relevanţă este diminuată de calitatea sistemului de traducere automată.

În sfârşit, o a treia variantă de rezolvare a problemei de interogare într-o limbă şi de căutare a răspunsului în documente elaborate în altă limbă este de a traduce automat doar grupurile sintactice (nu toată întrebarea) considerate relevante în construcţia frazei de interogare. Acest lucru revine la a traduce fraza de interogare construită pentru limba sursă. La ICIA au fost deja experimente în acest sens, calitatea traducerilor grupurilor nominale din engleză în română şi invers fiind extrem de bună. Una din direcţiile de dezvoltare a sistemului de întrebare-răspuns al ICIA, pe lângă transpunerea sa într-un serviciu web public, este încorporarea acestei abordări pentru tratarea situaţiilor interlinguale de întrebare-răspuns.

2.3.2 Eurovoc

Eurovoc este un tezaur multilingv (Steinberger et al., 2002), folosit la indexarea Acquis-ului comunitar. El conţine termeni din domeniile de activitate din ţările membre UE şi relaţii între aceştia. Eurovoc versiunea 4.3 există în 23 de limbi, acoperă 21 de domenii şi este organizat în 127 de microtezaure.

Termenii sunt de două feluri: descriptori (termenii folosiţi în indexare) şi non-descriptori (cuvinte din limba comună echivalente cu descriptorii, dar nefolosite în indexare; sunt folosite la expandarea frazei de interogare). Relaţiile semantice sunt de cinci tipuri: cele care definesc descriptorii, cele care indică apartenenţa acestora la unul sau mai multe microtezaure, relaţii de echivalenţă (stabilite între descriptori şi non-descriptori), relaţii ierarhice (stabilite între descriptori) şi asociative (reprezentând modalităţile în care descriptorii se pot asocia: cauză, efect, locaţie etc.)

Fiecare document din Acquis-ul comunitar este indexat cu termenii Eurovoc pe care îi conţine. Acest lucru facilitează găsirea răspunsului, prin limitarea căutării la acele documente care conţin descriptorii din întrebarea utilizatorului sau non-descriptorii aflaţi în diverse relaţii cu aceştia.

Sistemul de întrebare-răspuns dezvoltat la ICIA (Ion et al., 2008) face apel la Eurovoc în segmentarea şi lematizarea întrebării utilizatorului şi a

corpusului. Termenii din Eurovoc sunt consideraţi unităţi minimale în etapa de segmentare, iar la lematizare li se ataşează lema din Eurovoc[3].

2.4 Corpusul paralel Acquis comunitar

Nevoia de corpusuri paralele pentru aplicaţiile de procesare a limbajului natural a cunoscut o accentuare pe parcursul ultimilor ani. Corpusurile paralele sunt folosite în aplicaţiile de traducere automată sau categorizare multilingvală, pentru a produce resurse lexicale sau semantice multilinguale, cum sunt dicţionarele sau ontologiile lexicale, pentru a testa consistenţa procesului de traducere etc.

Pentru aplicaţiile de procesare a limbii române în context multilingual cel mai important corpus este JRC-Acquis (Steinberger et al., 2006). Acesta este, în prezent, cel mai mare corpus multilingual disponibil, conţinând 22 de limbi. Corpusul este disponibil în format XML conform specificaţiilor TEI (Text Encoding Initiative). De asemenea, conţine şi alinierea celor peste 231 de perechi de limbi conţinute de JRC-Acquis. Corpusul creşte pe măsura traducerii legislaţiei europene şi în limbile noilor candidaţi.

Acquis-ul comunitar este corpul comun de convenţii, legi şi obligativităţi care leagă toate statele membre ale Comunităţii Europene. Conţine principii şi obiective politice ale diverselor tratate semnate în cadrul Uniunii Europene, legislaţie UE, declaraţii şi rezoluţii, acorduri internaţionale şi obiective comune. Pe lângă cele 22 de limbi ale Comunităţii Europene, Acquis-ul este tradus şi în limbile croată şi turcă. Datorită efortului depus la ICIA pentru colectarea şi adnotarea documentelor Acquis-ului românesc, limba română a fost prezentă în pachetul de distribuţie JRC-Acquis încă de la prima versiune a acestuia.

Documentele disponibile în limba română sunt 19.211. Numărul de documente comune în perechea de limbi engleză-română este de 11.469. Documentele comune constituie un important corpus paralel conţinând 59.986.838 de cuvinte.

---

[3] În alegerea sinseturilor de implementat în Ro-WN atenţie deosebită a fost acordată acoperirii integrale a tezaurului Eurovoc.

Fişierele româneşti şi bulgăreşti nu sunt disponibile în acelaşi format (HTML) ca documentele celorlalte limbi din Acquis, neputând fi procesate de aceleaşi instrumente de convertire HTML-TEI. Fişierele în format HTML disponibile pentru celelalte limbi ale JRC-Acquis conţin şi informaţii cu privire la structura documentului, cum ar fi secţiunile de anexe şi semnături, secţiunile cu textul şi titlul documentului etc. Această structură nu se regăseşte în formatul Microsoft Word, în care sunt disponibile documentele româneşti ale JRC-Acquis.

Pentru a constitui colecţia de documente în limba română, fişierele au fost descărcate de pe site-ul „CCVista Translation Database" (http://ccvista.taiex.be/Fulcrum/CCVista/RO/<celex>, unde <celex> este numărul unic de identificare al documentului). Numărul total de documente în limba română disponibile pe site-ul CCVista este de 19.286.

Fişierele au fost convertite din formatul Microsoft Word în formatul XML conform specificaţiilor TEI. Conversia celor 19.286 fişiere a fost făcută automat, folosind pachetul de funcţii „Visual Studio Tools for Office". Acestea permit interacţiunea cu aplicaţia Microsoft Office direct din mediul de programare.

Datorită particularităţilor formatului, conversia documentelor a implicat şi o serie de etape intermediare: au fost înlăturate comentariile traducătorilor, au fost şterse notele de subsol şi secţiunile de cap de pagină, a fost normalizată folosirea caracterelor diacritice (unele documente foloseau „ş" şi „ţ" cu sedilă, iar altele „ş" şi „ţ" cu virgulă).

Dintre cele 19.286 de fişiere în format Microsoft Word au fost convertite 19.211 (restul de documente având erori de format). În formatul TEI-XML al documentelor româneşti au fost adăugate şi datele de indexare Eurovoc acolo unde acestea erau disponibile. În (Ceauşu, 2008) este descris pe larg procesul de colectare şi procesare a documentelor româneşti ale corpusului JRC-Acquis.

3 Concluzii

Aplicaţiile din lingvistica computaţională, indiferent de tip, nu pot avea performanţe competitive în absenţa unor resurse lingvistice de foarte bună calitate. Aceste aplicaţii reprezintă, de cele mai multe ori, o combinare de module separate, fiecare cu propriile necesităţi informaţionale. De exemplu, un sistem de întrebare-răspuns necesită o prelucrare primară a întrebării

utilizatorului pentru a putea extrage automat din aceasta o frază de interogare pentru motorul de căutare. La rândul ei, prelucrarea primară constă în adnotare morfolexicală şi lematizare, procese care nu pot funcționa fără existența unor lexicoane sau modele de limbă, acestea din urmă extrase din corpusuri adnotate morfolexical. În consecință, resursele lingvistice computaționale reprezintă, în cele din urmă, o aproximare (materializată prin elaborarea de modele de limbă, lexicoane, ontologii lexicale etc.) a competențelor lingvistice ale vorbitorilor.

Calitatea resurselor lingvistice computaționale este direct responsabilă de performanțele algoritmilor de prelucrare automată a limbajului natural. Resursele prezentate în această lucrare au fost folosite de ICIA la competiția sistemelor de întrebare-răspuns pentru limba română CLEF 2009 în care Institutul nostru a ocupat locul întâi. Se confirmă faptul că aceste resurse lingvistice computaționale sunt valoroase şi că orice efort de a le perfecționa reprezintă o foarte bună investiție din toate punctele de vedere.

Mulțumiri

Cercetările şi rezultatele prezentate în această lucrare au fost susținute prin proiectul nr. D11-007 „SIR-RESDEC” din programul național de cercetare PN2.

Referințe

Bentivogli, L., Forner, P., Magnini, B., Pianta, E. *Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing*. În Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources", Geneva, Switzerland, August 28, 2004, p. 101-108.

Bloksma, L., Diez-Orzas, Vossen, P., *The User Requirements and Functional Specification of the EuroWordNet-project*. EWN-deliverable D.001, LE-4003, 1996.

Ceauşu, A. *Colectarea şi procesarea documentelor româneşti ale corpusului JRC-Acquis*. În Resurse lingvistice şi instrumente pentru prelucrarea limbii române, Iaşi, 2008, p. 125-130.

Ceauşu, A., Ştefănescu, D., Tufiş, D. *Acquis Communautaire sentence alignment using Support Vector Machines*. În Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, 2006.

Chen, Q., Li, M., Zhou, M. *Improving Query Spelling Correction Using Web Search Results*. În Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, 2007, p. 181-189.